

Achtung!

Dies ist eine Internet-Sonderausgabe des Aufsatzes
„Das Projekt eines Thesaurus des Türkischen.
Computergestützte Untersuchungen zum türkischen Lexikon“
von Jost Gippert (1993).
Sie sollte nicht zitiert werden. Zitate sind der Originalausgabe in
„Laut- und Wortgeschichte der Türksprachen“,
hrsg. v. B. Kellner-Heinkele und M. Stachowski,
Wiesbaden: Harrassowitz 1995, 59-74
zu entnehmen.

Attention!

This is a special internet edition of the article
“Das Projekt eines Thesaurus des Türkischen.
Computergestützte Untersuchungen zum türkischen Lexikon”
[“The project of a Turkic thesaurus.
Computer-aided investigations into Turkic lexicography”]
by Jost Gippert (1993).
It should not be quoted as such. For quotations, please refer to the
original edition in
“Laut- und Wortgeschichte der Türksprachen”,
hrsg. v. B. Kellner-Heinkele und M. Stachowski,
Wiesbaden: Harrassowitz 1995, 59-74.

Alle Rechte vorbehalten / All rights reserved:

Jost Gippert, Frankfurt 2011

DAS PROJEKT EINES THESAURUS DES TÜRKISCHEN
COMPUTERGESTÜTZTE UNTERSUCHUNGEN ZUM TÜRKISCHEN LEXIKON

Das hier vorzustellende Projekt, das in baldiger Zukunft an der Universität Bamberg realisiert werden soll, ist der Erstellung eines computergestützten Thesaurus der türkischen Sprache im 20. Jh. gewidmet. Der Thesaurus soll die Entwicklung des Wortschatzes in der türkischen Schriftsprache seit der Umstellung auf die lateinschriftliche Graphie dokumentieren, wobei Etappen von jeweils einer Dekade angesetzt werden. Als Grundlage sollen einerseits typische Erzeugnisse der schöngeistigen Literatur (v.a. Prosa) dienen, andererseits verschiedene Bereiche der Zeitungssprache (Wirtschaft, Politik). Die betreffenden Texte sollen jeweils vollständig ausgewertet werden, um als Basis für statistische Analysen und eine sich daraus ergebende Stratifikation des verwendeten Wortschatzes zu dienen. Als Ergebnis sollen einerseits Spezialwörterbücher (z.B. zur Wirtschaftssprache) entstehen, die sich durch die Angabe von Belegstellen auf die zugrunde gelegten Texte beziehen; andererseits soll das gesammelte Wortmaterial im Hinblick auf seine Repräsentation in den Handwörterbüchern des Türkischen überprüft und ein diese supplierendes Lexikon erarbeitet werden. Der Thesaurus soll seiner Anlage nach für die zukünftige Sprachentwicklung offen bleiben; dazu soll er der Fachwelt auch in elektronischer Form zur Verfügung gestellt werden. Hinzu kommen Einsatzmöglichkeiten in der Lehre, die den Einsatz elektronischer Medien voraussetzen.

Das Projekt soll im wesentlichen vier Arbeitsabschnitte umfassen. Der erste Schritt besteht in der Vorbereitung der Textanalyse durch eine Umwandlung der zu analysierenden Texte in eine elektronische Form. Diese Umwandlung kann, dem heutigen Stand der Technik entsprechend, weitgehend automatisiert, nämlich unter Zuhilfenahme von optischen Scannern erfolgen. Dabei wird ein System benötigt, das auf die Besonderheiten der türkischen Graphie, insbesondere die in der elektronischen Verarbeitung als "Sonderzeichen" geltenden Buchstaben *ı, İ, ş, Ş, ğ, Ğ*, eingestellt werden kann und darüber hinaus eine Abgleichung der gelesenen Wortformen mit einem vorgegebenen Wortvorrat erlaubt (sog. "spell checking").

Die Encodierung muß ihrerseits wie folgt vorbereitet werden: Zunächst wird ein Grundvorrat von Wortformen gesammelt und in eine lexikonähnliche Liste umgearbeitet, die als Grundlage für das "Spell-Checking" dienen soll. Dazu kann auf den "Elementarwortschatz Türkisch-Deutsch" von Nuran TEZCAN (Wiesbaden 1988) zurückgegriffen werden, der bereits in elektronischer Form verfügbar ist. Da hieraus im wesentlichen aber nur Lexikonformen hervorgehen (Nominative, Infinitive etc.) und eine automatische morphologische Analyse und Zurückführung aller denkbaren Flexionsformen auf die Lexikoneinträge während des Spell-

Checkings nicht ohne weiteres möglich ist, müssen die zu jedem Eintrag bildbaren Einzelformen (Dativ Pl., 3.Ps.Sg.Futur o.ä.) eigens erzeugt werden. Dies läßt sich im Falle des Türkischen mit seiner relativ regelmäßigen Morphologie mit Hilfe eines Computerprogramms erzielen, dessen Entwicklung bereits abgeschlossen ist. Mit Hilfe des so erstellten "Basislexikons" lassen sich allfällige Lesefehler beim Scanning, die v.a. durch Ungenauigkeiten beim Druck entstehen, bereits auf ein geringes Maß reduzieren.

Der zweite Arbeitsvorgang betrifft die Analyse der eingescannten Textmaterialien im Hinblick auf den in ihnen repräsentierten Wortschatz. Da das Augenmerk hierbei auf die Dokumentation von Spezifika gerichtet ist, kann dieser Vorgang nur teilweise automatisiert ablaufen. Das einfachste Verfahren würde darin bestehen, zunächst den gesamten Wortvorrat des jeweiligen Textes zu extrahieren, was mit Hilfe existierender Computerprogramme schnell und problemlos möglich ist, und diesen Wortvorrat dann gegen den wie oben definierten "Elementarwortschatz" abzugleichen und so die "spezifischen" Wortformen herauszufiltern. Dieses Verfahren kann jedoch nicht für sich allein angewendet werden, da es unter die "spezifischen" Wortformen auch erhalten gebliebene Fehlesungen des Scanners oder Druckfehler subsumieren würde. Stattdessen soll ein Verfahren gewählt werden, das gleichzeitig eine Extraktion "spezifischer" Wortformen **und** eine endgültige Korrektur des eingescannten Textes ermöglicht; dieser Vorgang soll prinzipiell interaktiv, d.h. unter ständiger Steuerung durch Bearbeiter ablaufen. Benötigt wird eine Routine, bei der ein gegebener Text Wort für Wort mit dem vorliegenden Basislexikon verglichen wird und die bei allen nicht vorgefundenen Wörtern bzw. Wortformen eine Korrektur oder die Aufnahme in ein Supplementlexikon erlaubt; an einer solchen Routine wird derzeit gearbeitet.

Der dritte und zugleich aufwendigste Teilabschnitt des hier vorgestellten Projekts umfaßt die Auswertung der bei der Textanalyse gewonnenen Daten im Hinblick auf ihre zeitliche, textspezifische oder sonstige Zugehörigkeit und darauf aufbauend die Erstellung von Speziallexika.

Grundlage dieser Auswertung sollen die Supplementlexika bilden, die im Zuge des zweiten Projektvorgangs erstellt werden. Diese Supplementlexika werden naturgemäß Wortformen aller morphologischer Kategorien enthalten, so wie sie in den betreffenden Texten auftreten. Der erste Arbeitsvorgang muß folglich darin bestehen, die einzelnen Wortformen unter ihre jeweiligen Lexikonformen zu ordnen, da für eine lexikographische Auswertung primär letztere interessant sind. Dieser Vorgang kann wiederum weitgehend automatisiert werden; erforderlich ist die Entwicklung eines Computerprogramms, das die jeweiligen Wortformen auf der Basis der türkischen Wortbildungs- und Flexionsregeln zu analysieren versucht und lediglich in Zweifelsfällen einen interaktiven Eingriff des Bearbeiters verlangt; ein solches Programm kann prinzipiell dieselben Routinen verwenden, wie sie für das "Spell-Checking" erarbeitet werden.

Die mit Hilfe des zu entwickelnden Computerprogramms "lemmatisierten" Wortlisten sollen dann zunächst als Basis für eine Indizierung der eingegebenen Texte dienen. Bei dieser Indizierung ergeben sich zum einen die statistischen Daten, die über eine chronologische o.a. Zugehörigkeit der einzelnen Wörter Auskunft geben; zum anderen bildet sie die Grundlage für Belegstellenangaben, mit denen die avisierten Speziallexika versehen werden sollen.

Einen eigenen Auswertungsvorgang, der nur zum geringen Teil automatisiert werden kann, bildet die Überprüfung der eingelesenen Textmaterialien auf idiomatische Wendungen hin, die sich aus zwei oder mehr Wörtern zusammensetzen. Hierfür ist zunächst die Ausarbeitung einer Textstellenkonkordanz erforderlich, die jedes Wort in einem genügend großen Kontext, etwa Satz oder Teilsatz sichtbar macht. Die Feststellung idiomatischer Wendungen muß dann durch den Bearbeiter allerdings "von Hand" durchgeführt werden, und sie wird der zeitaufwendigste Schritt sein; der Bearbeiter muß hierfür über muttersprachliche Kenntnisse des Türkischen verfügen.

Der vierte Projektabschnitt betrifft die Abgleichung der gewonnenen lexikographischen Informationen mit den großen Handwörterbüchern des Türkischen, dem Türkçe Sözlük, Steuerwald und Redhouse. Um das gewonnene Material auf seine Repräsentation in diesen Wörterbüchern hin überprüfen zu können, soll deren Inhalt Eintrag für Eintrag elektronisch gespeichert werden, wobei lediglich die Lemmata selbst bzw. damit gebildete idiomatische Wendungen benötigt werden. Diese Eingabe ist beim gegebenen Stand der Technik sinnvoll nur von Hand durchzuführen, da bei der Eingabe mit dem Scanner automatisch auch sämtliche Begleitangaben, Übersetzungen etc. mit eingelesen würden, die für den hier vorgesehenen Arbeitsvorgang jedoch unerheblich, wenn nicht sogar störend sind.

Nach erfolgter Eingabe kann zunächst automatisch ein Gesamtindex des in den drei Wörterbüchern enthaltenen Wortmaterials erstellt werden, mit dem dann sämtliche Einzelwortlisten der bearbeiteten Texte oder auch ein gegen Projektende zu generierender Gesamtindex des aus den Texten hervorgehenden Wortmaterials abgeglichen werden kann. Daraus resultiert ein Verzeichnis von Wörtern, die in keinem der Standardwörterbücher erfaßt sind; dieses Verzeichnis soll als "Supplementlexikon des Türkischen" für sich veröffentlicht werden.

Der gegenwärtige Arbeitsvorgang, über den hier berichtet werden soll, besteht in der Erstellung der benötigten Programmodule zur automatischen Erzeugung und Analyse türkischer Wortformen.

Wie oben bereits gesagt wurde, sind im Hinblick auf die Erzeugung türkischer Wortformen (auf der Grundlage von Lexikoneinträgen, d.h. Infinitiven bzw. Nominativen) und für eine Analyse beliebiger in Texten auftretender Wörter prinzipiell dieselben Vorgaben zu beachten, und es können zum großen Teil dieselben Programmelemente verwendet werden: Um den Computer anzuweisen, von einem Verb wie *acıkma* 'hungrig werden' eine 1.Person Singular Perfekt *acıktım* zu

bilden, müssen ihm dieselben Informationen gegeben werden, wie wenn er ein gegebenes *acıktım* als 2. Person Singular Perfekt von *acıkmak* "erkennen" soll. Diese Informationen lassen sich als eine rel. geringe Menge von Kategorialisierungen fassen, die gewissermaßen das Grundgerüst der türkischen Formenbildung darstellen und die z.Tl. automatisch aus der Lexikon- oder Grundform des Wortes ableitbar sind.

Zunächst ist eine generelle Scheidung zwischen Verben und Nomina erforderlich, da sich hieraus unterschiedliche Mengen bildbarer Formen ableiten. Bei den Verben ist es sinnvoll, statt von dem Infinitiv vom einfachen Stamm auszugehen, von dem der Infinitiv selbst eine Weiterbildung darstellt; als Stamm kann jeweils das vor dem Infinitivsuffix *-mak/-mek* erscheinende Element gelten. Vom so isolierten Stamm ausgehend lassen sich dann die folgenden kategorialen Einteilungen bestimmen:

a) *vordervokalisch / hintervokalisch*: Diese Unterscheidung ergibt sich regelmäßig und ohne Ausnahmen aus dem Vokal der letzten Stammsilbe (Beispiel: *başla-* 'anfangen' vs. *söyle-* 'sagen'); durch sie wird die Wahl zwischen äquivalenten Suffixpaaren wie *-sa-* / *-se-* gesteuert (vgl. *başla-sa* / *söyle-se*);

b) *gerundet / ungerundet*: Diese Dichotomie ergibt sich ebenfalls ausnahmslos aus dem Vokal der letzten Stammsilbe (Beispiele: *uç-* 'fliegen' vs. *aç-* 'öffnen' oder *çök-* 'sinken' vs. *çek-* 'ziehen'); im Zusammenspiel mit der Unterscheidung *vordervokalisch / hintervokalisch* steuert sie die Wahl zwischen Suffixen mit vier Varianten wie z.B. *-miş* / *-müş* / *-muş* / *-mü* (vgl. *çek-miş* / *çök-müş* / *aç-muş* / *uç-muş*);

c) *vokalisch auslautend / konsonantisch auslautend*: Diese Divergenz ist wegen dreierlei Regelungen bedeutsam: nach vokalisch auslautenden Stämmen verlieren gewisse Suffixe den anlautenden Vokal (z.B. *acı-* 'hungrig werden' mit Aorist *acı-ır* vs. *acı-* 'schmerzen' mit Aorist *acı-r* oder *dinle-* 'sich erholen' mit Aorist *dinle-ir* vs. *dinle-* 'zuhören' mit Aorist *dinle-r*), vor manchen wird ein *-y-* "eingeschoben" (z.B. *gel-* 'kommen' mit Futur *gel-ecek* vs. *söyle-* 'sagen' mit Futur *söyle-y-ecek*), und vor bestimmten Suffixen wird der stammauslautende Vokal selbst "getilgt" (z.B. *başla-* 'anfangen' mit Präsens *başl-ıyor* oder *söyle-* 'sagen' mit *söyl-iyor*). Auch diese Divergenz ist normalerweise am Stamm selbst ablesbar; es sind jedoch zwei Sonderfälle zu berücksichtigen, nämlich *de-mek* 'sagen' und *ye-mek* 'essen', bei denen zusätzlich zu dem einzuschubenden *-y-* ein Vokalwechsel im Stammauslaut auftritt (z.B. Futur *di-y-ecek*, *yi-y-ecek*)¹;

¹ Die hier wiedergegebenen Regelungen sind natürlich an der heutigen Orthographie orientiert. Die im Rahmen des Projekts zu bearbeitenden Textmaterialien werden darüber hinaus zweifellos auch andere (frühere) Regelungen repräsentieren, so daß z.B. Formen wie *söyli-yecek* oder *başla-yor* als mögliche Varianten erfaßt werden müssen.

d) *weitvokalisch / engvokalisch*: Diese beiden Kategorien betreffen die Wahl des "Bindevokals" bei der Aoristbildung; ob bei einem gegebenen Verbalstamm das Aoristsuffix "weitvokalisch" (in der Form *-ar-* bzw. *-er-*) oder "engvokalisch" (*-ir-* / *-ür-* / *-ır-* / *-ur-*) zu erscheinen hat, ist aus der Form des Stammes *nicht* vorhersagbar (vgl. z.B. *ak-* 'fließen' mit Aor. *ak-ar* vs. *al-* 'nehmen' mit *al-ır*, *bin-* 'einsteigen' mit Aor. *bin-er* vs. *bil-* 'kennen' mit *bil-ir*, *sön-* 'verlöschen' mit Aor. *sön-er* vs. *öl-* 'sterben' mit *öl-ür* oder *tut-* 'halten' mit Aor. *tut-ar* vs. *bul-* 'finden' mit *bul-ur*). Weitvokalische Aoriste kommen zwar hauptsächlich bei einsilbigen Verbalstämmen vor, sind aber nicht völlig auf diese beschränkt (vgl. z.B. *vazgeç-* 'verzichten' mit Aor. *vazgeç-er* oder *hallet-* 'lösen' mit Aor. *halled-er*); auch wenn es sich bei Fällen wie den letztgenannten ausschließlich um Kompositalstämme handelt, ist ihre Aoristbildung doch nicht ohne weiteres aus der Form des Stammes ableitbar.

e) *stimmlos / nicht stimmlos*: Diese Unterscheidung betrifft Stämme mit konsonantischem Auslaut; sie regelt zunächst die lautliche Gestalt von mit Verschußlaut anlautenden Suffixen, deren Anlaut an den Stammauslaut assimiliert wird (Beispiele: *kaç-* 'flüchten' mit Perfekt *kaç-tı* vs. *kal-* 'bleiben' mit *kal-dı* oder *tut-* 'halten' mit Perfekt *tut-tu* vs. *dur-* 'stehenbleiben' mit *dur-du*). Diese Regelung gilt wiederum ausnahmslos und geht aus der Form des Stammes selbst hervor; unvorhersagbar ist demgegenüber eine weitere kategoriale Divergenz, die mit ihr verknüpft ist, nämlich

f) *sonorisierend / nicht sonorisierend*: Bei gewissen Stämmen, die auf stimmlosen Okklusiv endigen, wird dieser vor antretenden vokalisch anlautenden Suffixen stimmhaft (Beispiele: *et-* 'machen' mit Aor. *ed-er* oder *git-* 'gehen' mit *gid-er* vs. *öt-* 'singen' mit Aor. *öt-er*). Diese Regelung ist wiederum nicht aus der Form des Stammes ableitbar. Es wäre natürlich verlockend, die Regelung gewissermaßen umzukehren und die hierunter fallenden Beispiele als Stämme auf stimmhaften Okklusiv aufzufassen, der vor bestimmten Suffixen stimmlos wird (also z.B. *ed-* mit Infinitiv *et-mek* oder Perfekt *et-ti* bzw. *gid-* mit Infinitiv *git-mek* oder Perfekt *git-ti*). Da es jedoch auch Stämme auf stabile stimmhafte Okklusive gibt (z.B. *yed-* 'hinter sich herziehen') mit Inf. *yed-mek* oder *eğ-* 'krümmen' mit Inf. *eğ-mek*), bestünde bei einer derartigen Notierung des Stammes keine Eindeutigkeit.

g) Als ein Sonderfall für sich bleibt das Verbum *i-* 'sein' bestehen, das bekanntlich über eine eingeschränkte Formenbildung verfügt.

Aus dem Gesagten dürfte bereits deutlich geworden sein, daß auch für die an die Verbalstämme tretenden Suffixe kategoriale Unterscheidungen sinnvoll sind. Zunächst lassen sich diese nach der Anzahl der vorliegenden lautlichen Varianten rubrizieren, d.h. ob sie z.B. zwei Varianten (wie das Konditionalsuffix *-sa-* vs. *-se-*), vier Varianten (wie das Suffix *-miş-* / *-müş-* / *-muş-* / *-muş-*), sieben Varianten (wie das Aoristsuffix mit den Formen *-r-* / *-er-* / *-ar-* / *-ir-* / *-ür-* / *-ır-*

/ -ur-) oder, als Maximum, acht Varianten kennen (wie das Perfektsuffix -di- / -dii- / -di- / -du- / -ti- / -tü- / -ti- / -tu-). Die Anzahl der Varianten hängt aber ihrerseits wieder von der lautlichen Struktur ab, so daß für jeden Verbalstamm, der mit Hilfe der o.g. Kategorien eindeutig definiert ist, die jeweilige zu benutzende Variante eindeutig vorhersagbar ist; benötigt wird dann jeweils nur eine Normalform ("Archimorphem"). So lassen sich z.B. folgende Regelungen beobachten:

a) nach dem Merkmal *vordervokalisch* / *hintervokalisch* bei einem gegebenen Verbalstamm wechseln im Suffix die Vokale -e- / -a- bzw. -i- / -ı-;

b) nach dem zusätzlichen Merkmal *gerundet* / *ungerundet* wechseln -i- / -ii- bzw. -ı- / -u- (während -e- und -a- keinem weiteren Wechsel unterliegen);

c) nach dem Merkmal *vokalisch auslautend* / *konsonantisch* auslautend wechselt das Suffix -r- zwischen dieser und einer um einen Vokal erweiterten Form (-er-, -ar-, -ir-, -ır-, -ür-, -ur-), ferner erhalten Suffixe, die mit Vokal anlauten, Varianten mit einem "eingeschobenen" -y- (z.B. -ecek- / -yecek-);

d) nach dem Merkmal *weitvokalisch* / *engvokalisch* wechselt der Vokal des Suffixes -r- (Aorist) bei konsonantisch auslautenden Stämmen zwischen -e- / -a- auf der einen und -i- / -ii- / -ı- / -u- auf der anderen Seite;

e) nach dem Merkmal *stimmlos* / *nicht stimmlos* wechseln mit Verschlußlaut anlautende Suffixe zwischen stimmlos und stimmhaft (z.B. -t- und -d- beim Perfektsuffix);

f) das Merkmal *sonorisierend* / *nicht sonorisierend* beim Verbalstamm betrifft die Suffixe nicht.

Ein Computerprogramm, das sich diese Regelmäßigkeiten zunutze machen soll, benötigt also rel. wenige Zusatzinformationen, um zu einem gegebenen Verbalstamm die korrekten Formen zu erzeugen: Lediglich die nicht aus der Form des Verbalstamms selbst ableitbare Entscheidung zwischen *weitvokalischen* und *engvokalischen* Aoristbildungen muß ihm zu jedem Verbalstamm eigens mitgeteilt werden, alle anderen Differentiationen können in das Programm selbst pauschal eingebaut werden (was auch das irreguläre Verb *i-* 'sein' betrifft). Das gleiche gilt auch für die Analyse von in Texten vorgefundenen Verbalformen: um eine Form wie *acıktı* 'er bekam Hunger' richtig zu analysieren, braucht dem Programm lediglich der Verbalstamm *acık-* vorgegeben zu sein, wodurch sich die Variante -tı- des in der Normalform z.B. als -di- ansetzbaren Perfektsuffixes automatisch ergibt, da sich der Verbalstamm selbst als *hintervokalisch*, *ungerundet*, *konsonantisch*, *stimmlos* offenbart.

Das bisher zum Verbum gesagte gilt natürlich nur für die Bildung von temporalen und modalen Kategorien. Als kaum vorhersagbar dürften hingegen die Möglichkeiten der Kausativ- oder Passivbildung gelten; hier fehlen nicht nur eindeutige formale Kriterien, sondern es ist vielfach auch rein semantisch begründet, ob eine solche abgeleitete Bildung überhaupt existiert oder existieren kann. Deshalb ist es ratsam, solche abgeleitete Bildungen zunächst als eigene Verben zu

fassen; als Pauschalregel mag dabei gelten, daß alles als eigenes Verb behandelt werden kann, was einen eigenen Infinitiv hat.

Ganz ähnliche, aber nicht völlig identische Regelungen sind auch bei Nominalstämmen zu berücksichtigen. Zunächst gibt es auch hier wieder eine Unterscheidung

a) *vordervokalisch* / *hintervokalisch*: Auch hier gilt prinzipiell die Regel, wonach der Vokal der letzten Stammsilbe die Vokalqualität der Suffixe bestimmt (vgl. z.B. *bay* 'Herr' mit Dat. *bay-a* vs. *bey* 'Herr' mit Dat. *bey-e*); allerdings gibt es hier Ausnahmen (z.B. *saat* 'Stunde' mit Dat. *saat-e*), die in keiner Weise aufgrund ihrer formalen Struktur vorhersagbar sind. Ob ein Nominalstamm vorder- oder hintervokalisch ist, muß also zusätzlich angegeben werden. – Ebenso wie bei den Verben gibt es weiter die Dichotomie

b) *gerundet* / *ungerundet*: Diese ist im Gegensatz zur obigen Unterscheidung eindeutig am Vokal der letzten Stammsilbe ablesbar (vgl. z.B. *bey* 'Herr' mit Akk. *bey-i* vs. *köy* 'Dorf' mit *köy-ü* oder *bay* 'Herr' mit Akk. *bay-ı* vs. *boy* 'Größe' mit *boy-u*). – Von grundlegender Bedeutung ist dann die Unterscheidung

c) *konsonantisch* / *vokalisch*: Sie hat entscheidenden Einfluß auf die Form sowohl mit Vokal anlautender als auch mit Konsonant anlautender Suffixe, wobei jeweils mehrere Unterkategorien zu beachten sind. Für konsonantisch auslautende Nominalstämme sind dies die Differentiationen

ca) *stimmlos* / *stimmhaft*: Sie regelt, wie bei Verbalstämmen, den Anlaut von Suffixen wie der Lokativ- oder Ablativendung, deren *-d-* an einen auslautenden stimmlosen Konsonanten assimiliert wird (vgl. z.B. *ay* 'Mond, Monat' mit Lok. *ay-da* vs. *kitap* 'Buch' mit *kitap-ta* oder *ev* 'Haus' mit Lok. *ev-de* vs. *sepet* 'Korb' mit *sepet-te*). Bei okklusivem Stammauslaut ist zusätzlich ebenso wie beim Verbum die folgende Unterscheidung zu beachten:

cb) *sonorisierend* / *nicht sonorisiert*: Ob ein Stamm auf Verschußlaut diesen vor vokalisch anlautenden Suffixen stimmhaft werden läßt (vgl. z.B. *ceket* 'Jacke' mit Akk. *ceket-i* vs. *çeşit* 'Sorte' mit *çeşid-i* oder *köpek* 'Hund' mit Akk. *köpeğ-i* vs. *yük* 'Last' mit *yük-ü*), läßt sich nicht vorhersagen, wenn man als Stamm die jeweilige Nominativform nimmt. Wie beim Verbum ist es ausgeschlossen, als Stamm jeweils die sonorisierte Variante (also *çeşid-*, *köpeğ-*) anzugeben, da Stämme auf stabile stimmhafte Okklusive existieren (z.B. *ad* 'Name', *dağ* 'Berg', *rab* 'Herr'). – Weitere Sonderfälle entstehen durch die Dichotomie

cc) *geminert* / *nicht geminiert*: Daß bei bestimmten Nominalstämmen, die auf einfachen Konsonanten endigen, dieser vor mit Vokal anlautenden Suffixen geminiert wird (vgl. z.B. *hat* 'Linie' mit Akk. *hatt-ı* vs. *at* 'Pferd' mit *at-ı* oder *hak* 'Recht' mit Akk. *hakk-ı* vs. *ok* 'Pfeil' mit *ok-u*, mit zusätzlicher Sonorisierung z.B. *zıt* 'Gegensatz' mit Akk. *zıdd-ı*), ist wiederum aus der Nominativform nicht ableitbar. Auch hier wäre es denkbar, den Stamm stattdessen in der Form zu

notieren, wie er vor der Akkusativendung erscheint (also *hatt*, *at*, *hakk*, *ok*, *ııdd*). – Ähnliches gilt letztlich für die Dichotomie

cd) *synkopierend* / *nicht synkopierend*: Vom Nominativ ausgehend ist nicht vorhersagbar, ob ein Wort den Vokal der letzten Silbe vor dem Antritt vokalisches anlautender Suffixe synkopiert oder nicht (vgl. *gümüş* 'Silber' mit Akk. *gümüş-ü* vs. *göğüs* 'Brust' mit *göğs-ü*). Denkbar wäre hier, den Stamm in der Form zu verzeichnen, die er im Akkusativ annimmt (also *gümüş-*, aber *göğs-*); da der der Synkope unterliegende Vokal den Regeln der Vokalharmonie unterliegt (also den oben aufgestellten Kategorien vordervokalisch / hintervokalisch und gerundet / ungerundet), wäre er in jedem Fall vorhersagbar (also *göğs-* > *göğüs* 'Brust', *ağz-* > *ağız* 'Mund', *oğl-* > *oğul* 'Sohn', *şehir-* > *şehir* 'Stadt')². Die nötige Eindeutigkeit ergibt sich weiter daraus, daß die bei der Synkope entstehenden Konsonantengruppen nicht in den Wortauslaut treten können (im Gegensatz zu "echten" Konsonantengruppen wie *-nk* in *renk* 'Farbe' oder *-rç* in *borç* 'Schuld').

Für *vokalisch* auslautende Nominalstämme gelten die folgenden Kategorialisierungen:

ce) als *normal* können solche Stämme auf Vokal gelten, bei denen, formalistisch gesprochen, vor den vokalisches anlautenden Endungen des Dativs und Akkusativs ein *-y-*, vor der Genetivendung ein *-n-* und vor der Possessivendung der 3. Person Singular ein *-s-* "eingeschoben" wird (vgl. z.B. *oda* 'Stube' mit Akk. *oda-yı* oder *gezi* 'Ausflug' mit Akk. *gezi-yi*). Dem stehen jedoch folgende Sonderfälle gegenüber:

cf) Stämme, bei denen kein Konsonanteneinschub erfolgt (z.B. *camî* 'Moschee' mit Akk. *camî-i* [neben *camî-yî*]); da es sich hierbei ursprünglich um Stämme auf arabisches ^cAyn handelt, könnten diese als *aynhaltig* bezeichnet werden. Sie sind aufgrund der Nominativform ebensowenig vorhersagbar wie

cg) Stämme, bei denen der stammauslautende Vokal vor vokalisches anlautenden Suffixen synkopiert (und durch einen Apostroph ersetzt) wird (z.B. *nevi* 'Art, Spezies' mit Akk. *nev'i*); auch hier liegt ein arabisches ^cAyn zugrunde, und man könnte zur Unterscheidung vom vorherigen Typ von *apostrophierend* sprechen. Historisch handelt es sich natürlich um eine Synkope wie bei den Stämmen unter cd), so daß man auch hier die Variante *nev'-* als Stamm einsetzen könnte; dies würde die Regel erfordern, daß der Apostroph (also das ursprüngliche ^cAyn) bei der Bildung des Nominativs, Lokativs usw. (*nevi*, *nevi-de*) durch ein *i* ersetzt werden muß (eine solche Regel muß natürlich eine Abgrenzung von Eigennamen des Typs *Ankara* mit Lok. *Ankara'da* gewährleisten). Einen weiteren Sonderfall unter vokalisches auslautenden Stämmen bilden

² Das gilt auch für Fälle wie *vakt-* > *vakit* 'Zeit', da solche Wörter ohnehin als vordervokalisch gekennzeichnet werden müssen (vgl. Akk. *vakt-i*).

ch) Stämme auf Vokal, an die die Akkusativendung in "possessivischer" Form, d.h. mit "eingeschobenem" *-n-* antritt (z.B. *demiryolu* 'Eisenbahn' mit Akk. *demiryolu-nu*). Hierbei handelt es sich meist um unverbundene Possessivausdrücke, so daß man sie als *possessivisch* bezeichnen kann (daneben allerdings auch Pronomina wie *bu* 'dieser' mit Akk. *bu-nu*). Zu beachten ist, daß sie wiederum an der Form des Nominativs nicht erkennbar sind (vgl. z.B. *basımevi* 'Druckerei' mit Akk. *basımevi-ni* vs. *gezi* 'Ausflug' mit Akk. *gezi-yi* oder *nevi* 'Art' mit Akk. *nev'-i*). Dasselbe gilt letztlich für das unregelmäßige Wort

ci) *su* 'Wasser' und seine Komposita (z.B. *akarsu* 'fließendes Wasser'), bei denen vor den Possessivendungen und der Genetivendung ein *-y-* eingeschoben wird (vgl. *su-yun* Gen. 'des Wassers' mit *şu-nun* Gen. 'dessen').

Wie beim Verbum, so fallen auch beim Nomen Kategorialisierungen der Suffixe an, die mit den hier aufgelisteten Regelungen korrelieren. Zunächst unterscheiden sie sich ebenfalls aufgrund der Anzahl der Varianten, die sie bilden; die Bandbreite erstreckt sich von zwei Varianten (z.B. beim Pluralsuffix *-lar-* / *-ler-*) über vier (z.B. bei der Lokativendung *-da* / *-de* / *-ta* / *-te*), sechs (bei der Dativendung *-a* / *-e* / *-ya* / *-ye* / *-na* / *-ne*), zwölf (bei der Akkusativendung *-ı* / *-u* / *-i* / *-ü* / *-yı* / *-yu* / *-yi* / *-yü* / *-ni* / *-nu* / *-ni* / *-nü*) bis zu dreizehn (beim Possessivsuffix der 3. Person Singular: *-ı* / *-u* / *-i* / *-ü* / *-sı* / *-su* / *-si* / *-sü* / *-ni* / *-nu* / *-ni* / *-nü* / *-yu*). Wie beim Verbum hängt die Wahl der jeweiligen Variante aber ihrerseits wieder von der lautlichen Struktur des vorausgehenden Stammes ab, so daß sie immer eindeutig vorhersagbar ist; benötigt wird auch hier nur ein "Archimorphem". Man vgl. z.B. die folgenden Regelungen:

a) nach dem Merkmal *vordervokalisch* / *hintervokalisch* bei einem gegebenen Nominalstamm wechseln im Suffix die Vokale *-e-* / *-a-* bzw. *-i-* / *-ı-*;

b) nach dem zusätzlichen Merkmal *gerundet* / *ungerundet* wechseln *-i-* / *-ü-* bzw. *-ı-* / *-u-* (während *-e-* und *-a-* keinem Wechsel unterliegen);

c) nach dem Merkmal *konsonantisch auslautend* / *vokalisch auslautend* besteht z.B. die Akkusativendung entweder aus einfachem Vokal oder ist um einen Konsonanten erweitert (*-i*, *-ı*, *-ü*, *-u* / *-yi*, *-yı*, *-yü*, *-yu*, *-ni* usw.);

d) aufgrund des Merkmals *aynhaltig* bei *vokalisch* auslautenden Stämmen fällt der "eingeschobene" Konsonant z.B. vor der Akkusativendung wieder aus (*-si* > *-i*);

e) nach dem Merkmal *stimmlos* / *nicht stimmlos* (bzw. *okklusiv* / *nicht okklusiv*) bei *konsonantischen* Stämmen wechseln mit Verschußlaut anlautende Suffixe zwischen stimmlos und stimmhaft (z.B. *-t-* und *-d-* bei der Ablativendung) usw.

Das avisierte Computerprogramm benötigt im Bereich des Nomens also einige Zusatzinformationen mehr als beim Verbum, wenn es zu jedem gegebenen Stamm die korrekten Formen erzeugen bzw. analysieren können soll: Schon die Differenzierung zwischen *vordervokalisch* und *hintervokalisch* ist aufgrund der

Stammform nicht mehr eindeutig möglich; bei *konsonantischem* Stammauslaut sind weiter *sonorisierende*, *geminierende* und *synkopierende* Stämme zu kennzeichnen, und bei *vokalischen* Stämmen *aynhaltige*, *apostrophierende*, *possessivische* sowie der Sonderfall *su*.

Bisher wurde nur über Verben und Nomina gesprochen. Natürlich gibt es daneben aber auch noch andere Wortarten, die eine eigene Behandlung erfordern. Dabei ist zu beachten, das sie sich äußerlich, etwa aufgrund der Lexikonform, in keiner Weise von Nominalstämmen zu unterscheiden brauchen; man vergleiche z.B. die Partikel *acaba* 'etwa' mit *araba* 'Wagen' oder das Adverb *ait* 'zu .. gehörig' mit *bit* 'Laus'. Für das avisierte Computerprogramm können diese Unterscheidungen nun in verschiedener Weise verarbeitet werden. Zum einen wäre es denkbar, die verschiedenen Wortarten als solche zu kodieren und dann im Programm selbst die Regelungen zu verankern, in welcher Weise sie jeweils abgewandelt werden können; dies würde z.B. implizieren, daß eine Information vorgegeben sein müßte, wonach von einem Adverb wie *ait* Prädikatsformen der 3. Person Singular (*ait-tir* 'es ist zu .. gehörig') sinnvoll bildbar sind, kaum jedoch von einer Fragepartikel (z.B. [†]*acaba-dir*)³.

Nun gibt es allerdings auch bei den verschiedenen Nominalbildungen durchaus systematische Beschränkungen in der Formenbildung, die zu berücksichtigen sind. So bilden z.B. die Pronomina *bazı*, *biri* und *kimi* keine Possessivformen der beiden ersten Personen Singular; der Stamm *üzer-* tritt zwar in Formen wie *üzer-ine*, *üzer-inde* auf, nicht jedoch allein (quasi als Nominativ) usw.

Noch problematischer wird die Frage der bildbaren Formen, wenn man semantische Aspekte berücksichtigt. So ist z.B. fraglich, ob ein Wort wie *millet* 'Nation' sinnvoll Prädikatsformen der beiden ersten Personen Singular bilden kann (**millet-im* 'ich bin die/eine Nation' ?) oder ob es einen Kontext gibt, wo eine Form **çocuğ-umuz-um* 'ich bin unser Kind' ihren Platz hätte.

Diese Fragen sind für das angestrebte Computerprogramm sowohl dort relevant, wo es darum geht, einen Vorrat von korrekten türkischen Wortformen zu *erzeugen*, als auch im Hinblick auf die *Analyse* gegebener Textmaterialien, da diese Analyse ja auch auf eine Korrektur eventueller Lesefehler abzielen soll und somit helfen soll, hinter einem eventuell als *acabadır* gelesenen Wort das wahrscheinlichere *arabadır* zu vermuten. Deshalb erscheint es zweckmäßig, die Formenbildung zunächst überhaupt restriktiv zu handhaben, d.h. Formen, die als systematischen oder semantischen Gründen zweifelhaft sind, zunächst einmal nicht durch das Programm generieren zu lassen. Auf eine Unterscheidung etwa

³ Von metasprachlichen Verwendungen wie *acaba-dir* 'es ist, sc. das Wort, *acaba*', auf die mich M. ERDAL anlässlich meines Vortrags hinwies, kann dabei zunächst abgesehen werden.

zwischen Nomina, Pronomina, Adverbien etc. kann dabei sogar verzichtet werden, da diese aus den nötigen Zusatzangaben über bildbare Formen ohnehin hervorgeht; stattdessen brauchen nur noch Verben und Nicht-Verben einander gegenübergestellt zu werden.

Im Zusammenhang mit den Kategorialisierungen der Wortstämme ist bereits die Frage angeklungen, wie die Daten für das Computerprogramm am günstigsten aufbereitet werden können, d.h., welche Codierung zu wählen ist. Das übergeordnete Prinzip dürfte dabei in einer möglichst platzsparenden Notation bestehen; denn es ist ja beabsichtigt, das Programm für Benutzer normaler PCs mit ihrem begrenzten Arbeitsspeicher verfügbar zu machen. Dabei ist zu bedenken, daß allein für finite Verbalformen im Türkischen rund 500 Endungsprototypen existieren, für Nominalformen sogar rund 1500, wenn man alle denkbaren Kombinationen von Possessivsuffixen, Kasusendungen und Prädikatsformen berücksichtigt (vgl. die Ausschnitte in den Tafeln I und II).

Es versteht sich von selbst, daß es sinnlos wäre, das Programm einfach mit einer (vielleicht alphabetisch sortierten) Liste aller auf diese Weise bildbaren Wortformen operieren zu lassen⁴, zumal der Vorrat ja beliebig erweiterbar sein soll (um Wörter bzw. Formen, die nicht zum Grundwortschatz gehören, sondern erst in den zu analysierenden Texten begegnen). Stattdessen sollen Stämme und Suffixe getrennt verwaltet werden, wobei es darauf ankommt, die den oben ausgeführten grammatischen Regelungen entsprechenden Kategorialisierungen ebenfalls möglichst platzsparend zu codieren.

⁴ Diesbezügliche Überlegungen (allerdings im Hinblick auf maschinelle Übersetzung) bereits bei Vietze-Zenker-Warnke, Rückläufiges Wörterbuch der türkischen Sprache, Leipzig 1975, 6.

Nominale Formenbildung (Ausschnitt)				
<i>kadınımız</i>	<i>kadınımızın</i>	<i>kadınımıza</i>	<i>kadınımızda</i>	<i>kadınımızdan</i>
? <i>kadınımızım</i>	? <i>kadınımızımın</i>	? <i>kadınımızayım</i>	<i>kadınımızdayım</i>	<i>kadınımızdanım</i>
<i>kadınımızsın</i>	? <i>kadınımızsının</i>	? <i>kadınımızasın</i>	<i>kadınımızdasın</i>	<i>kadınımızdansın</i>
<i>kadınımızdır</i>	<i>kadınımızındır</i>	<i>kadınımızadır</i>	<i>kadınımızdadır</i>	<i>kadınımızdandır</i>
? <i>kadınımızız</i>	? <i>kadınımızınız</i>	? <i>kadınımızayız</i>	<i>kadınımızdayız</i>	<i>kadınımızdanız</i>
<i>kadınımızsınız</i>	? <i>kadınımızsınızın</i>	? <i>kadınımızasınız</i>	<i>kadınımızdasınız</i>	<i>kadınımızdansınız</i>
? <i>kadınımızdırlar</i>	<i>kadınımızındırlar</i>	<i>kadınımızadırlar</i>	<i>kadınımızdadırlar</i>	<i>kadınımızdandırlar</i>
? <i>kadınımızdım</i>	? <i>kadınımızdımın</i>	? <i>kadınımızaydım</i>	<i>kadınımızdaydım</i>	<i>kadınımızdandım</i>
<i>kadınımızdın</i>	? <i>kadınımızdının</i>	? <i>kadınımızaydın</i>	<i>kadınımızdaydın</i>	<i>kadınımızdandın</i>
<i>kadınımızdı</i>	<i>kadınımızındı</i>	<i>kadınımızaydı</i>	<i>kadınımızdaydı</i>	<i>kadınımızdandı</i>
? <i>kadınımızdık</i>	? <i>kadınımızdıkın</i>	? <i>kadınımızaydık</i>	<i>kadınımızdaydık</i>	<i>kadınımızdandık</i>
<i>kadınımızdınız</i>	? <i>kadınımızdınızın</i>	? <i>kadınımızaydınız</i>	<i>kadınımızdaydınız</i>	<i>kadınımızdandınız</i>
? <i>kadınımızdırlar</i>	<i>kadınımızındırlar</i>	<i>kadınımızadırlar</i>	<i>kadınımızdadırlar</i>	<i>kadınımızdandırlar</i>
? <i>kadınımızmışım</i>	? <i>kadınımızmışımın</i>	? <i>kadınımızaymışım</i>	<i>kadınımızdaymışım</i>	<i>kadınımızdanmışım</i>
<i>kadınımızmışsın</i>	? <i>kadınımızmışsının</i>	? <i>kadınımızaymışsın</i>	<i>kadınımızdaymışsın</i>	<i>kadınımızdanmışsın</i>
<i>kadınımızmış</i>	<i>kadınımızmışın</i>	<i>kadınımızaymış</i>	<i>kadınımızdaymış</i>	<i>kadınımızdanmış</i>
? <i>kadınımızmışız</i>	? <i>kadınımızmışınızın</i>	? <i>kadınımızaymışız</i>	<i>kadınımızdaymışız</i>	<i>kadınımızdanmışız</i>
<i>kadınımızmışsınız</i>	? <i>kadınımızmışsınızın</i>	? <i>kadınımızaymışsınız</i>	<i>kadınımızdaymışsınız</i>	<i>kadınımızdanmışsınız</i>
? <i>kadınımızmışlar</i>	<i>kadınımızmışların</i>	<i>kadınımızaymışlar</i>	<i>kadınımızdaymışlar</i>	<i>kadınımızdanmışlar</i>
? <i>kadınımızsam</i>	? <i>kadınımızsamın</i>	? <i>kadınımızaysam</i>	<i>kadınımızdaysam</i>	<i>kadınımızdansam</i>
<i>kadınımızsan</i>	? <i>kadınımızsanın</i>	? <i>kadınımızaysan</i>	<i>kadınımızdaysan</i>	<i>kadınımızdansan</i>
<i>kadınımızsa</i>	<i>kadınımızsanın</i>	<i>kadınımızaysa</i>	<i>kadınımızdaysa</i>	<i>kadınımızdansa</i>
? <i>kadınımızsak</i>	? <i>kadınımızsakın</i>	? <i>kadınımızaysak</i>	<i>kadınımızdaysak</i>	<i>kadınımızdansak</i>
<i>kadınımızsanız</i>	? <i>kadınımızsanızın</i>	? <i>kadınımızaysanız</i>	<i>kadınımızdaysanız</i>	<i>kadınımızdansanız</i>
<i>kadınımızsalar</i>	<i>kadınımızsaların</i>	<i>kadınımızaysalar</i>	<i>kadınımızdaysalar</i>	<i>kadınımızdansalar</i>

Tafel I

(?: fragwürdige Formen)

Was zunächst die Codierung der Stämme betrifft, so ist es sinnvoll, die Zusatzangaben in computergerechter Weise in kleinstmögliche Informationseinheiten, sog. Bits, zu "verpacken". Dabei lassen sich die wichtigsten der nach dem bisher gesagten nötigen Informationen durchaus in einem einzigen Zeichen mit acht Bits (d.h. einem Byte) pro Stamm unterbringen. Wenn man davon ausgeht, daß sich in jedem Bit eine Entscheidung zwischen den zwei Zuständen "ja" und "nein" niederschlägt, so läßt sich das z.B. in Form eines Baumdiagramms darstellen, wo jeder Knoten ein Bit und seine Abzweigungen nach links bzw. rechts den positiven bzw. negativen Zustand darstellen (s. Abb. 1). Dabei fällt auf, daß die vorgegebenen acht Entscheidungen bei verbalen Stämmen nur z.T. genutzt sind, bei nominalen Stämmen hingegen noch weitere Unterscheidungen wünschenswert wären.

Verbale Formenbildung (Ausschnitt)

acıkın	acıkiyormuşsun	acıkiyordurdur	acıktıldır	acıkırlarsaymış
acıktınız	acıkiyormuş	acıkiyorsunuzdur	acıktıysam	acıksana
acıksın	acıkiyormuşuz	acıkiyorlardır	acıktımsa	acıksanıza
acıksınlar	acıkiyormuşsunuz	acıkmıyorumdur	acıktıysan	acıkma
acıkiyorum	acıkiyorlarmış	acıkmıyorsundur	acıktınsa	acıkmayın
acıkiyorsunuz	acıkiyormuşlar	acıkmıyordur	acıktıysa	acıkmayınız
acıkiyor	acıkmıyormuşum	acıkmıyordurdur	acıktıysak	acıkmasana
acıkiyoruz	acıkmıyormuşsun	acıkmıyorsunuzdur	acıktıksa	acıkmasanıza
acıkiyorsunuz	acıkmıyormuş	acıkmıyorlardır	acıktıysanız	acıkmasın
acıkiyorlar	acıkmıyormuşuz	acıkmışındır	acıktınızsa	acıkmasınlar
acıkmıyorum	acıkmıyormuşsunuz	acıkmışındır	acıktılarsa	acıkmadım
acıkmıyorsunuz	acıkmıyorlarmış	acıkmıştır	acıktıysalar	acıkmadın
acıkmıyor	acıkmıyormuşlar	acıkmıştıdır	acıktırım	acıkmadı
acıkmıyorsunuz	acıkmışmışım	acıkmışsındır	acıkırsın	acıkmadık
acıkmıyorlar	acıkmışmışınız	acıkmışlardır	acıktır	acıkmadınız
acıkmışım	acıkmışmış	acıkmıştırlar	acıkırtı	acıkmadılar
acıkmışsın	acıkmışmışız	acıkiyormuşumdur	acıkırsınız	acıkmamışım
acıkmış	acıkmışmışınız	acıkiyormuşsundur	acıkırlar	acıkmamışsın
acıkmışız	acıkmışmışlar	acıkiyormuştur	acıkırdım	acıkmamış
acıkmışsınız	acıkmışlarmış	acıkiyormuşsundur	acıkırdın	acıkmamışız
acıkmışlar	acıkmışsam	acıkiyormuşsunuzdur	acıkırdı	acıkmamışsınız
acıkiyordum	acıkmışsan	acıkiyorlarmıştır	acıkırdık	acıkmamışlar
acıkiyordun	acıkmışsa	acıkiyormuşlardır	acıkırdınız	acıkacağım
acıkiyorduk	acıkmışsak	acıkmıyormuşumdur	acıkırlardı	acıkacaksın
acıkiyordunuz	acıkmışsanız	acıkmıyormuşsundur	acıkırdılar	acıkacak
acıkiyorlardı	acıkmışlarsa	acıkmıyormuştur	acıkırmışım	acıkacağız
acıkmıyordum	acıkiyorsam	acıkmıyormuşsundur	acıkırmışsın	acıkacaksınız
acıkmıyordun	acıkiyorsan	acıkmıyormuşsunuzdur	acıkırmış	acıkacaklar
acıkmıyorduk	acıkiyorsa	acıkmıyorlarmıştır	acıkırmışız	acıkmayacağım
acıkmıyordunuz	acıkiyorsa	acıkmıyormuşlardır	acıkırmışsınız	acıkmayacaksın
acıkmıyorlardı	acıkiyorsak	acıktım	acıkırlarmış	acıkmayacak
acıkmıyorlar	acıkiyorsalar	acıktın	acıkırsam	acıkmayacağız
acıkmıyorlardı	acıkmıyorsam	acıktı	acıkırsan	acıkmayacaksınız
acıkmıyorlardı	acıkmıyorsan	acıktık	acıkırsa	acıkmayacaklar
acıkmıştım	acıkmıyorsa	acıktınız	acıkırsak	acıkmam
acıkmıştın	acıkmıyorsak	acıktılar	acıkırsanız	acıkmazsın
acıkmıştı	acıkmıyorsanız	acıktımdı	acıkırlarsa	acıkmaz
acıkmıştık	acıkmıyorlarsa	acıktındı	acıkırsammış	acıkmayız
acıkmıştınız	acıkmıyorsalar	acıktıydı	acıkırsanmış	acıkmazsınız
acıkmışlardı	acıkiyordumdur	acıktıktı	acıkırsaymış	acıkmazlar
acıkmıştırlar	acıkiyorsundur	acıktınızdı	acıkırsakmış	
acıkiyormuşum	acıkiyordur	acıktıydılar	acıkırsanızmış	usw.

Deshalb kann zunächst daran gedacht werden, bei den Verben noch zusätzliche Informationen unterzubringen, die z.B. die wortbildungsmäßige Stellung eines Verbalstamms betreffen; so könnten Markierungen für Kausativa, Passiva, reziproke Bildungen u.ä. vorgesehen werden, die im Hinblick auf die durch das Programm zu leistenden Auswertungen bedeutsam werden können. Denkbar ist darüber hinaus, generell noch ein zweites Informationsbyte (also weitere acht Bits = Einzelinformationen) vorzusehen; dieses könnte bei den nicht verbalen Stämmen die fehlenden Subkategorialisierungen sowie evtl. Wortbildungsinformationen aufnehmen, bei Verben zusätzliche Angaben zur Rektion, die v.a. im Hinblick auf idiomatische Wendungen interessant sein dürften. An der günstigsten und zugleich aussagekräftigsten Codierung wird derzeit noch gearbeitet.

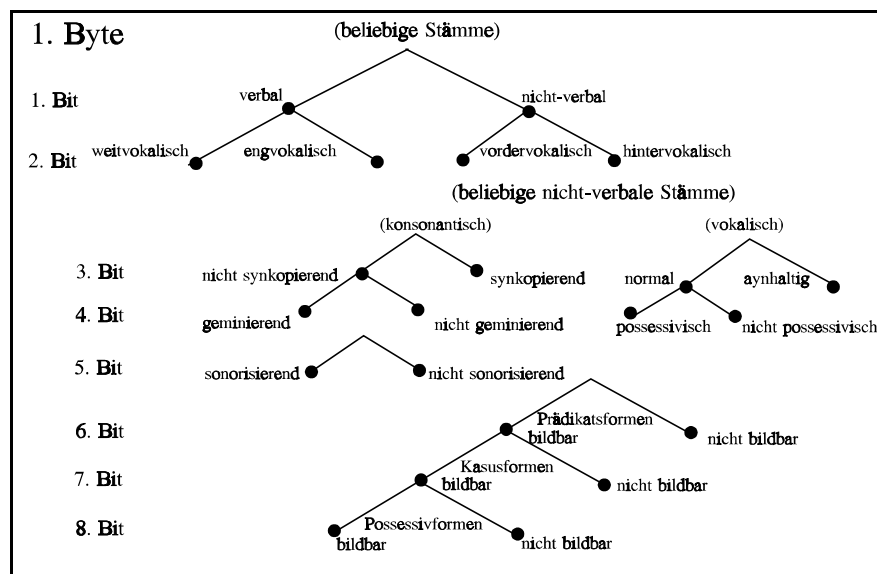


Abb. 1

Bei der Codierung der Suffixverbände sind zwei prinzipiell unterschiedliche Verfahren denkbar, ein analytisches und ein synthetisches.

Das analytische Verfahren würde darin bestehen, daß jedes Suffixelement für sich gespeichert würde (wiederum als "Archimorphem", d.h. in einer "unmarkierten" Form) und das Programm die Regeln über die Variations- und Kombinations-

möglichkeiten enthalten würde; ein solches Verfahren kann für das Nomen etwa durch folgende Aufstellung verdeutlicht werden:

<i>ler</i>	<i>im</i>		<i>dim</i>
	<i>in</i>	<i>in</i>	<i>din</i>
	<i>i</i>	<i>e</i>	<i>di</i>
	<i>imiz</i>	<i>i</i>	<i>dik</i>
	<i>iniz</i>	<i>de</i>	<i>diniz</i>
	<i>leri</i>	<i>den</i>	<i>dilar</i>

Eine derartige Codierung hätte den Vorteil, daß die Gesamtmenge der zu berücksichtigenden Suffixe gut überschaubar bliebe; außerdem würde das Programm so als Nebeneffekt auch statistische Auswertungen im Hinblick auf die einzelnen Morpheme gestatten⁵. Dem stünden jedoch dreierlei Nachteile gegenüber: Zum einen sind die lautlichen Vorgänge bei der Kombinatorik komplex; so wären innerhalb der Kombinationen wieder ganz ähnliche Kategorialisierungen anzuwenden wie an der Grenze zwischen Stamm und Suffix (man vgl. z.B. die Variation beim Perfektsuffix in den Verbänden *-iyor-dum*, *-muş-tum* und *-se-ydim*). Zum zweiten unterliegen die Kombinationsmöglichkeiten nicht immer systematischen Regelungen hinsichtlich ihrer Anordnung: so gibt es z.B. die Kombination *-imiz-den-dir* (Possessivsuffix - Ablativendung - Prädikatsendung 3.Ps.Sg.), aber nicht *†-den-dir-imiz*, *†-dir-den-imiz* usw., es gibt *-miş-tir* (Inferentialsuffix – Prädikatsendung 3.Ps.Sg.), aber nicht *†-dir-miş*, eine Verdoppelung von *-ler-* (Pluralsuffix) ist ausgeschlossen (also *-ler-i* statt *†-ler-leri*, nominales Pluralsuffix – Possessivsuffix 3.Ps.Pl.) usw. Zum dritten sind gewisse Kombinationen, wie oben bereits angedeutet, aus semantischen Gründen fragwürdig oder ausgeschlossen; so ist z.B. *-im-dir* geläufig (Possessivsuffix 1.Ps.Sg. – Prädikatsendung 3.Ps.Sg.), während *-im-iz* kaum auftreten dürfte (Possessivsuffix 1.Ps.Sg. – Prädikatsendung 1.Ps.Pl.; das entsprechende gilt auch für die 2. Person), und die Kombination einer Prädikatsendung mit der Lokativendung kann als normal gelten (z.B. *-de-yim*), während die Kombination mit der Akkusativendung ausgeschlossen erscheint (*†-i-yim*). Diese Beschränkungen in das Programm selbst einzubauen, bedeutet nicht nur einen rel. hohen Programmieraufwand, sondern wird sich in jedem Fall auch auf die bei jeder Einzelanalyse anfallende Rechenzeit auswirken.

⁵ Vgl. die diesbezüglichen Untersuchungen von J.E. PIERCE in *Anthropological Linguistics* 3, 1961, 31-42 und 4, 1962, 30-41.

Bei einer "synthetischen" Codierung würde demgegenüber jeder vorkommende Suffixverband als ganzer gespeichert, wobei es wiederum genügen würde, jeweils eine Variante als "Archimorphem" zu verwenden. Für das Verbalssystem käme so z.B. eine Liste wie die in Tafel III wiedergegebene zustande.

Auch dieses Verfahren hätte einen Nachteil, nämlich daß die Gesamtmenge der zu verwaltenden Suffixverbände außerordentlich groß ist; es handelt sich, wie gesagt, um ca. 500 Prototypen für Verbalendungen und ca. 1500 für Nomina (inkl. Prädikatsformen). Dieser Nachteil würde jedoch dadurch aufgewogen, daß die Regeln für die interne Kombinatorik, nachdem die Liste einmal erzeugt ist, weitgehend ignoriert werden können und die Regeln für die externe Kombinatorik, d.h. die Anbindung an gegebene Wortstämme, aufgrund der für diese vorzugebenden Kategorialisierungen rel. leicht zu fassen sind. Insgesamt scheint das "synthetische" Verfahren daher den Vorzug zu verdienen.

<i>iyorum</i>	<i>iyordum</i>	<i>iyormuşum</i>	<i>mişim</i>	<i>mişmişim</i>
<i>iyorsun</i>	<i>iyordun</i>	<i>iyormuşsun</i>	<i>mişsin</i>	<i>mişmişsin</i>
<i>iyor</i>	<i>iyordu</i>	<i>iyormuş</i>	<i>miş</i>	<i>mişmiş</i>
<i>iyoruz</i>	<i>iyorduk</i>	<i>iyormuşuz</i>	<i>mişiz</i>	<i>mişmişiz</i>
<i>iyorsunuz</i>	<i>iyordunuz</i>	<i>iyormuşsunuz</i>	<i>mişsiniz</i>	<i>mişmişsiniz</i>
<i>iyorlar</i>	<i>iyordular</i>	<i>iyormuşlar</i>	<i>mişler</i>	<i>mişmişler</i>
<i>miyorum</i>	<i>miyordum</i>	<i>miyormuşlar</i>	<i>miştin</i>	<i>mişlermiş</i>
<i>miyorsun</i>	<i>miyordun</i>	<i>miyormuşum</i>	<i>miştin</i>	<i>mişsem</i>
<i>miyor</i>	<i>miyordu</i>	<i>miyormuşsun</i>	<i>mişti</i>	<i>mişsen</i>
<i>miyoruz</i>	<i>miyorduk</i>	<i>miyormuş</i>	<i>miştik</i>	<i>mişse</i>
<i>miyorsunuz</i>	<i>miyordunuz</i>	<i>miyormuşuz</i>	<i>miştiniz</i>	<i>mişsek</i>
<i>miyorlar</i>	<i>miyorlardi</i>	<i>miyormuşsunuz</i>	<i>mişlerdi</i>	<i>mişseniz</i>
		<i>miyorlarmış</i>	<i>miştiler</i>	<i>mişlerse</i>
		<i>miyormuşlar</i>		usw.

Tafel III

Um die Informationsmenge, die mit Hilfe des Analyseprogramms gewonnen werden kann, von vorneherein möglichst groß zu halten, ist es erforderlich, auch bei der Wahl des "synthetischen" Verfahrens jeden einzelnen Suffixverband mit zusätzlichen Angaben zu versehen, durch die seine grammatische Funktion erkennbar wird. Eine solche zusätzliche Codierung kann zugleich die Lemmatisierung der aus den zu untersuchenden Texten gewonnenen Wortindizes erleichtern, indem sie eine automatische Reihung der Wortformen (z.B. *kadın*, *kadının*, *kadını*, *kadına*, *kadında*, *kadından*, d.h. Nominativ, Genetiv, Akkusativ, Dativ, Lokativ, Ablativ bei den Kasusformen) zu erzielen gestattet, die von der "computergerechten" rein alphabetischen Anordnung (vgl. z.B. *kadın*, *kadına*, *kadında*, *kadından*, *kadını*, *kadının*) abweicht. Auch an dieser Codierung wird derzeit noch gearbeitet.